# A Study on Mutual Information-based Feature Selection for Text Categorization

Yan Xu[1,2], Gareth Jones[3], JinTao Li[1], Bin Wang[1], ChunMing Sun[1,2]

[1] *Institute of Computing Technology,Chinese Academy of Sciences, Beijing,100080, China*
[2]*North China Electric Power University, Beijing,102206, China*
[3] *Dublin City University, Ireland*

**Abstract**

Feature selection plays an important role in text categorization. Automatic feature selection methods such as document frequency thresholding (DF), information gain (IG), mutual information (MI), and so on are commonly applied in text categorization. Many existing experiments show IG is one of the most effective methods, by contrast, MI has been demonstrated to have relatively poor performance. According to one existing MI method, the mutual information of a category c and a term t can be negative, which is in conflict with the definition of MI derived from information theory where it is always non-negative. We show that the form of MI used in TC is not derived correctly from information theory. There are two different MI based feature selection criteria which are referred to as MI in the TC literature. Actually, one of them should correctly be termed "pointwise mutual information" (PMI). In this paper, we clarify the terminological confusion surrounding the notion of "mutual information" in TC, and detail an MI method derived correctly from information theory. Experiments with the Reuters-21578 collection and OHSUMED collection show that the corrected MI method's performance is similar to that of IG, and it is considerably better than PMI.

*Keywords:* Text Categorization; Feature selection; Mutual Information; information theory

## 1. Introduction

Text categorization (TC) is the process of grouping texts into one or more predefined categories based on their content. Due to the increased availability of documents in digital form and the rapid growth of online information, text categorization has become one of the key techniques for handling and organizing text data.

Feature selection is an important step in TC, in recent years a growing number of statistical classification methods and machine learning techniques have been applied for this task. The prevailing feature selection methods include document frequency (DF) thresholding, information gain (IG), and mutual information (MI).

Most of the published results on feature selection show that the MI method has much lower performance than IG for text categorization [2][3][4].

But what is the "mutual information" measure used for feature selection in TC? According to [1][4][5][6][17] if a category $c$ and a term $t$, have probabilities $P(t)$ and $P(c)$, then their mutual information, $I(t,c)$, is defined to be:

$$I(t,c)=\log_2\frac{p(t,c)}{p(t)\times p(c)}=\log_2\frac{p(t\wedge c)}{p(t)\times p(c)} \qquad (1)$$

Informally, MI compares the probability of observing $t$ and $c$ together (the joint probability) with the probabilities of observing $t$ and $c$ independently (chance). If there is a genuine association between $t$ and $c$, then the joint probability $P(t,c)$ will be much larger than chance $P(t)\,P(c)$, and consequently $I(t,c) >> 0$. If there is no significance relationship between $t$ and $c$, then $P(t,c) \approx P(x)\,P(y)$, and thus, $I(t,c) \approx 0$. If $t$ and $c$ are in complementary distribution, then $P(t,c)$ will be much less than $P(x)\,P(y)$, forcing $I(t,c) << O$.

According to Equation 1, the MI of $t$ and $c$ can be negative, which is in conflict with the definition of MI in information theory where it is always non-negative, so it would seem that the mutual information used in [1][4][5][6][17] is not the one defined in information theory.

For example, let $p(t)$=0.8, $p(c)$=0.7, $P(t/c)$=0.5, then

$$I(t,c)=\log_2 \frac{p(t \wedge c)}{p(t)\times p(c)}=\log_2 \frac{0.5}{0.7\times 0.8}=\log_2 0.89<0$$

In information theory, the term "mutual information" refers to two random variables. It seems that (mostly in corpus-linguistic studies) that the term “mutual information” has been used for something which should correctly be termed "pointwise mutual information" as it is applied not to two random variables, but rather to two particular events from the sample spaces on which the two random variables are defined. This is the version used in current studies, and Equation 1 is really pointwise mutual information (PMI). Thus the "mutual information" method used for feature selection in TC should correctly be termed "pointwise mutual information".

In this paper, we describe a MI method derived from the original definition of information theory, we refer to this original mutual information method as the MI method. Experiments on the Reuters-21578 collection and OHSUMED collection show that the performance of the MI method is similar to that of IG, and is notably better than that of the PMI method.

The remainder of this paper is structured as follows: Section 2 describes the term selection methods, especially the PMI method, Section 3 describes the original MI method, Section 4 presents our experiments and results, and finally Section 5 summarizes our conclusions.

## 2. Feature Selection Methods

In this section we summarize and reexamine the feature selection methods DF, IG and PMI, commonly used in text categorization. DF, and IG both have good performance for feature selection in TC, with IG being generally superior [2][3][4].

The following definitions of DF, IG and PMI are taken from [3] and [4].

### *2.1. Document Frequency thresholding*

Document frequency is the number of documents in which a term occurs. Only the terms that occur in a large number of documents are retained. Yang and Pedersen’s experiments showed that it is possible to reduce the dimensionality by a factor of 10 with no loss in effectiveness [4][5].

DF thresholding is the simplest technique for vocabulary reduction. It scales easily to very large corpora with an approximately linear computational complexity in the number of training documents.

### *2.2. Information Gain*

Information gain is commonly used as a term goodness criterion in machine learning [7][8]. It measures the amount of information obtained for category prediction by knowing the presence or absence of a term in a document. Let m $\{c_i\}_{i=1}^m$ denote the set of categories in the target space. The information gain of term *t* is defined to be:

$$G(t) = -\sum_{i=1}^{m} p_r(c_i)\log p_r(c_i) + p_r(t)\sum_{i=1}^{m} p_r(c_i \mid t)\log p_r(c_i \mid t) + p_r(\bar{t})\sum_{i=1}^{m} p_r(c_i \mid \bar{t})\log p_r(c_i \mid \bar{t})$$

Given a training corpus, the information gain is computed for each unique term. Those terms whose information gain is less than some predetermined threshold are removed from the feature space.

### *2.3. Pointwise Mutual Information*

"Mutual information" is a criterion commonly used in statistical language modeling of word associations and related applications [5][6][9][17]. It should correctly be termed "pointwise mutual information" as it is not being applied to two random variables, since in information theory, the term "mutual information" refers to two random variables[10].

Given a category *c* and a term *t*, let *A* denote the number of times *c* and *t* co-occur, *B* denotes the number of times *t* occurs without *c*, *C* denotes the number of times *c* occur without *t*, and *N* denotes the total number of documents in *c*. The pointwise mutual information criterion between *t* and *c* is defined as:

$$PI(t,c) = \log \frac{p\ (t \wedge c)}{p\ (t) \times p\ (c)}$$

and is estimated using:

$$PI(t,c) \approx \log \frac{A \times N}{(A+C) \times (A+B)}$$

These category-specific scores of a term are then combined to measure the goodness of the term at a global level. Let $\{c_i\}_{i=1}^{m}$ denote the set of categories in the target space. Typically it can be calculated in one of two ways:

$$PI_{avg}(t) = \sum_{i=1}^{m} p\ (c_i) I(t,c_i), \qquad PI_{\max}(t) = \max_{i=1}^{m} \{I(t,c_i)\}$$

After the computation of these criteria, thresholding is performed to achieve the desired degree of feature elimination from the full vocabulary of a document corpus.

According to [6][17], in a general way, pointwise mutual information as defined above compares the probability of observing *t* and *c* together (the joint probability) with the probabilities of observing *t* and *c* independently (chance). If there is a genuine association between t and c, then the joint probability *P(t,c)* will be much larger than chance *P(t) P(c)*, and consequently *PI(t,c) >> 0*. If there is no significant relationship between *t* and *c*, then *P(t,c) ≈P(t)P(c)*, and thus, *PI(t,c) ≈0*. If t and c are in complementary distribution, then *P(t,c)* will be much less than *P(t) P(c)*, forcing *Pl(t,c) << O*. That is, pointwise mutual information as defined above can be negative, $PI_{avg}(t)$ also can be negative, in our experiments, $PI_{avg}(t)$ is found be negative for about 20% of the terms.

According to information theory, the MI of any random variables *X* and *Y* is always non-negative, so the pointwise mutual information as defined above is not actually the "mutual information" as defined in information theory.

### 3. Information Theoretic Mutual Information

In standard information theory research, the MI between two discrete random variables X and Y is defined to be [10][11]:

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} P(x,y) \log_2 \left( \frac{P(x,y)}{p(x)p(y)} \right) \qquad (2)$$

Additional properties are

$I(X;Y) = I(Y;X)$,

$I(X;Y) = H(X) + H(Y) - H(X;Y)$, and

(Non-negativity of mutual information): $I(X;Y) \geq 0$ with equality if and only if *X* and *Y* are independent.

Where $H(X)$ is the entropy of the random variable X. $H(Y)$ is the entropy of the random variable Y and $H(X;Y)$ is the joint entropy of these variables.

These category-specific term scores are then combined to measure the goodness of the term *t* at a global level. Let $\{c_i\}_{i=1}^{m}$ denote the set of categories in the target space, $C = \bigcup_{i=1}^{m} c_i$, let $T = \{t, \bar{t}\}$ denote the set in which term *t* occurs or *t* does not occur. According to (2), the mutual information criterion between T and C is defined as:

$$I\ (T;C) = \sum_{t' \in T} \sum_{c_i \in C} P(t', c_i) \log_2 \frac{p\ (t', c_i)}{p\ (t')p\ (c_i)}$$

$$= \sum_{i=1}^{m} P(t \wedge c_i) \log_2 \frac{p\ (t \wedge c_i)}{p\ (t) \times p\ (c_i)} + \sum_{i=1}^{m} P(\bar{t} \wedge c_i) \log_2 \frac{p\ (\bar{t} \wedge c_i)}{p\ (\bar{t}) \times p\ (c_i)}$$

$I\ (T;C)$ is the MI criterion between T and C,

According to the MI properties, $I\ (T;C) \geq 0,$

Let $I\ (t) = I\ (T;C)$

$$I(t) = \sum_{i=1}^{m} P(t \wedge c_i) \log_2 \frac{p\ (t \wedge c_i)}{p\ (t) \times p\ (c_i)} + \sum_{i=1}^{m} P(\bar{t} \wedge c_i) \log_2 \frac{p\ (\bar{t} \wedge c_i)}{p\ (\bar{t}) \times p\ (c_i)} \qquad (3)$$

So, $I\ (t) \geq 0$

Given a training corpus, for each unique term *t* we can compute the MI given by (3) and then remove from the feature space those terms whose information gain is less than some predetermined threshold, this is the MI method.

## 4. Experiment Results

Our objective is to compare the DF, IG and PMI methods with the MI method.

A number of statistical classification and machine learning techniques have been applied to text categorization, we use two different classifiers, the k-nearest-neighbor classifier (kNN) and the Naïve Bayes classifier. We chose kNN because evaluations have shown that it outperforms nearly all the other systems [12], and we selected Naïve Bayes because it is also one of the most efficient and effective inductive learning algorithm for classification [13].

Micro-averaging precision is widely used in cross-method comparisons [14], and we adopt it here to evaluate the performance of the different feature selection methods.

### *4.1. Data Collections*

Two corpora were used in our experiments: the Reuters-21578 [15] and the OHSUMED collection [16].

The Reuters-21578 collection is the original Reuters-22173 with 595 exact duplicste-documents removed, and has become a benchmark lately in text categorization evaluations.

OHSUMED is a bibliographical document collection. The documents were manually indexed using subject categories in the National Library of Medicine. There are about 1800 categories defined in MeSH, and 14321 categories present in the OHSUMED document collection.

### *4.2. Results*

Figures 1 and 2 show the performance curves of kNN and Naïve Bayes on the Reuters-21578 collection after feature selection using DF, IG, PMI, and MI. It can be seen in Figures 1 and 2 that the MI method outperforms the PMI method.

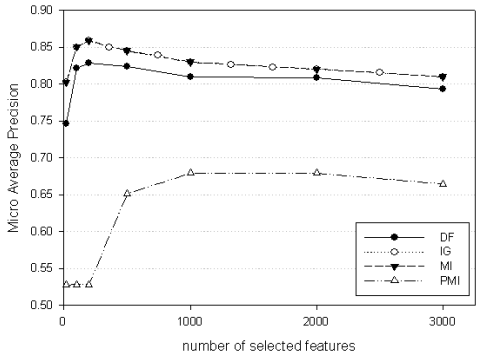


**Fig. 1.** Average precision of KNN vs. Number of selected features in Reuters.

.

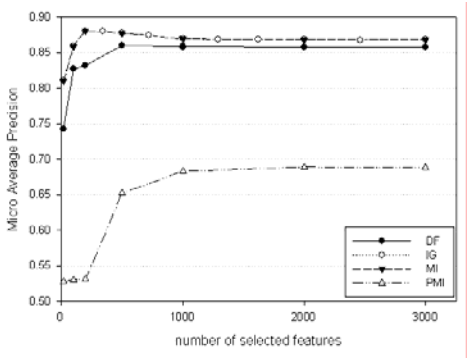


**Fig. 2.** Average precision of Naïve Bayes vs. Number of selected features in Reuters.

A further observation emerges from the categorization results of the kNN and the Naïve Bayes on Reuters. We find that IG and MI are the most effective in our experiments, that is, IG and MI produce similar performance of the classifiers. DF thresholding performed similarly. In contrast, PMI has by far the lowest performance.

Figures 3 and 4 show the performance curves of kNN and Naïve Bayes on OHSUMED after feature selection using DF, IG, PMI and MI. We observe the same result as that seen in Figures 1 and 2 where IG and MI are the most effective methods, DF performed similarly, in contrast, PMI had by far the poorest performance.

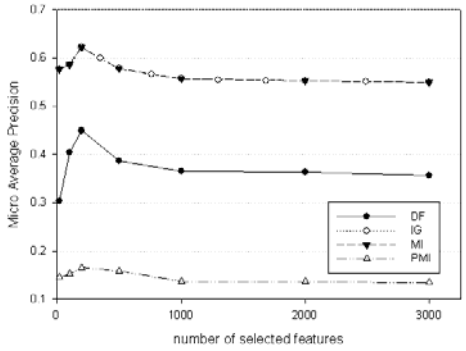


Fig. 3. Average precision of KNN vs. Number of selected features on OHSUMED.

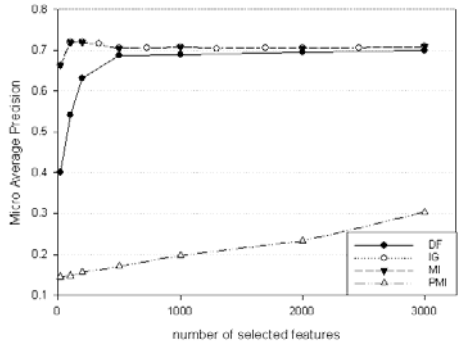


Fig. 4. Average precision of Naïve Bayes vs. Number of selected features on OHSUMED.

## 5. Conclusion

Information Retrieval has successfully brought together leading researchers and developers from many different areas. But sometimes, for the same concept, different areas have different nomenclature, the other

way round, for a dissimilar concept, different areas can have the same nomenclature.

In information theory, the term "mutual information" refers to two random variables. It has happened (mostly in corpus-linguistic studies) that this term has been used for something which should correctly be termed "pointwise mutual information" as it is applied not to two random variables, but rather to two particular events from the sample spaces on which the two random variables are defined.

In this paper:

- We point out this confusion in order to bring some sense into the terminological ambiguity surrounding the notion of "mutual information".
- We detail the mutual information (MI) method which is derived from the information theory.
- Experiments on the Reuters-21578 collection and the OHSUMED collection show that the MI method's performance is similar to that of IG, and it is observably better than PMI.

**Acknowledgement**

This work is supported by The National Natural Science Fundamental Research Project of China (60473002) and by The National Natural Science Fundamental Research Project of Beijing (4051004).

**References**

[1] S. Doan and S. Horiguchi, "An Efficient Feature Selection using Multi-Criteria in Text Categorization for Naive Bayes Classifier", WSEAS Transactions on Information Science and Applications, Vol.2, Issue 2, pp.98-103, 2005

[2] Y. Liu, A Comparative Study on Feature Selection Methods for Drug Discovery, J. Chem. Inf. Comput. Sci. 2004, 44, 1823-1828

[3] St. M.Yang, X.-B. Wu, Z.-H. Deng, M. Zhang and D.-Q. Yang. 2002 Modification of Feature Selection Methods Using Relative Term Frequency. Proceedings of ICMLC-2002, pp. 1432-1436

[4] Y. Yang, J. and O. Pedersen. 1997. A Comparative Study on Feature Selection in Text Categorization. Proceedings of ICML-97, pp. 412-420.

[5] F. Sebastiani, Machine Learning in Automated Text Categorization. ACM Computing Surveys, 34(1):1-47. 2002.

[6] K. W. Church and P. Hanks. Word Association Norms, Mutual Information and Lexicography. In Proceedings of ACL 27, pp.76-83, Vancouver, Canada,1989

[7] J.R. Quinlan. Induction of Decision Trees. Machine Learning, 1(1): pp.81-106, 1986

[8] T. Mitchell. Machine Learning. McCraw Hill, 1996

[9] E. Wiener, J.O. Pedersen, and A.S. Weigend. A Neural Network Approach to Topic Spotting. In Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval(SDAIR'95), 1995

[10] T. M. Cover and J. A. Thomas, Elements of Information Theory, 1991 John Wiley & Sons, Inc. Print ISBN 0-471-06259-6 Online ISBN 0-471-20061-1

[11] C. E. Shannon, ``A mathematical theory of communication," Bell System Technical Journal, vol. 27, pp. 379-423 and 623-656, July and October, 1948.

[12] Y. Yang and X. Liu. A re-examination of text categorization methods. (SIGIR'99), pp. 42-49, 1999

[13] H. Zhang. The Optimality of Naive Bayes. The 17th International FLAIRS conference, Miami Beach, May 17-19, 2004.

[14] Y. Yang. An Evaluation of Statistical Approaches to Text Categorization. Journal of Information Retrieval, Vol 1, No. 1/2, pp 67–88, 1999

[15] Reuters21578. http://www.daviddlewis.com/resources/testcollections/reuters21578/

[16] OHSUMED.http://www.cs.umn.edu/%CB%9Chan/data/tmdata.tar.gz

[17] R. Fano. 1961 Transmission of Information. MIT Press,Cambridge,MA